



**UNITED STATES DEPARTMENT OF COMMERCE**  
**Economics and Statistics Administration**  
**U.S. Census Bureau**  
Washington, DC 20233-0001

April 19, 2021

MEMORANDUM FOR

John M. Abowd  
Associate Director for Research and Methodology

From:

William R. Bell *Signed April 19, 2021*  
Senior Mathematical Statistician for Small Area Estimation  
Research and Methodology Directorate

Joseph L. Schafer *Signed April 19, 2021*  
Senior Mathematical Statistician for Estimation  
Research and Methodology Directorate

Subject:

Simulation Studies to Investigate Variation in Census Counts and  
in Census Coverage Error Using 2010 SF-1 Data and 2010 CCM  
Results

Attached are two notes documenting the methods and results from two studies we performed that used simulations to investigate variation in census counts (Joseph Schafer) and variation in census coverage error (William Bell), making use of SF-1 data from the 2010 census and results from the 2010 Census Coverage Measurement.

Attachments:

1. Block-Level Simulation of Non-Sampling Variability in Decennial Census Population Counts (Joseph Schafer)
2. Simulating Block-Level Populations Using 2010 Census Data and Coverage Measurement Results (William Bell)

cc: Patrick Cantwell (DSSD)  
Timothy Kennel (DSSD)  
Tom Mule (DSSD)  
Andrew Keller (DSSD)  
Scott Konicki (DSSD)

U S C E N S U S B U R E A U

# Block-Level Simulation of Non-Sampling Variability in Decennial Census Population Counts

Joseph L. Schafer\*

April 19, 2021

## Abstract

Population counts from a decennial census contain no sampling error, but they do reflect noise from non-sampling sources such as omissions, erroneous enumerations, mislocations, response errors, nonresponse and imputation of missing values. The unpredictable and uncontrolled nature of those processes implies that, if the same data collection and processing methods were repeatedly applied to the same fixed population, the resulting census counts would still vary. Using published block-level counts from the 2010 Census and results from the 2010 Census Coverage Measurement program, we performed an experiment to simulate the natural variability in population counts over repeated realizations of the census. Under conservative assumptions about error processes, the average deviation in state-level population counts was less than 0.1% of the population for every state. At the county level, the average deviation was  $\pm 117$  persons, or 0.3% of the county population. Across all census blocks with at least one housing unit, the average deviation was  $\pm 1.5$  persons. Because of the conservative assumptions applied in this simulation, these estimates should be interpreted as lower bounds on the natural variability in census population counts.

---

\*Office of the Associate Director for Research and Methodology, United States Census Bureau, Washington, DC 20233, [joseph.l.schafer@census.gov](mailto:joseph.l.schafer@census.gov).

# 1 Introduction

After each Decennial Census of Population and Housing, the United States Census Bureau publishes counts of persons in every state, followed by detailed figures for smaller regions, the smallest being census blocks. State-level totals are used for apportioning seats in the U.S. House of Representatives, and the more detailed numbers, which are provided in the *Census Redistricting Data (Public Law 94-171) Summary File*, are needed by states to delineate their congressional districts. Census data collection is performed on the entire population without statistical sampling, so these published counts contain no sampling variability. They are, however, subject to many types of non-sampling error, including but not limited to those listed below.

- **Omissions:** Persons who were eligible to be counted in the census but were not. Omissions may occur when entire housing units are missed, i.e., not listed on the Master Address File (MAF). Within units that are listed on the MAF, individuals may still be missed due to insufficient knowledge on the part of the respondent, misunderstanding or misapplication of residency rules, and for many other reasons.
- **Erroneous enumerations:** Persons reflected in the census count who were not eligible to be included because they were duplicated (already correctly counted at the same or a different location), nonexistent (fictitious persons), not alive or not residing in the United States on Census Day.
- **Mislocations:** Eligible persons who were counted but assigned to places outside the blocks where they resided on Census Day. An entire household will be mislocated if a dwelling is placed in the wrong block on the MAF. Within housing units that are correctly located, individuals may be mislocated for many reasons, e.g., students living away at college who are mistakenly counted at their parents' homes. Mislocations do not affect the population total for the nation, but they do impact how the population is distributed across subnational areas.
- **Response errors in characteristics.** Mistakes made when recording or processing demographic characteristics (age, sex, etc.) of persons can affect tabulations involving those characteristics.
- **Nonresponse and imputation of counts and missing characteristics.** When all attempts to obtain information have been exhausted, the occupancy status or number of persons living in some housing units remains unknown. Numbers of persons in these unresolved units are predicted by statistical procedures known as count imputation, and unknown characteristics for these persons are filled in by methods called characteristic imputation.

During and immediately after each census, the U.S. Census Bureau studies non-sampling errors to estimate their impact on data quality and to plan improvements for the next decade. Many of these projects fall under the Census Coverage Measurement (CCM) program and Census Program for Evaluation and Experiments (CPEX). Despite these efforts, of course, non-sampling errors can never be fully eliminated, and thus any published figure from a census will reflect some degree of systematic bias and random variation.

In this report, we focus on the following question: How much variation in population counts would be seen if a census could be conducted over and over, i.e., if the same general procedures for data collection and processing were repeatedly applied under similar conditions, with the underlying population held fixed?

Strictly speaking, the answer to this question is unknowable. Each census is a unique, unrepeatable event; by the time it has finished, the population has changed and the prevailing conditions have been altered by many happenings including the census itself. Nevertheless, by gathering evidence from data quality studies and published census figures, we can represent major components of non-sampling error with statistical models. In constructing these models, we estimate components of variation across domains within a census, and then use these as proxies for unobservable variation over hypothetical repetitions of the census. Studies of this type have a long history and have led to major changes and improvements in decennial census methodology (Hansen et al., 1961; Fellegi, 1964; Tepping and Bailer, 1973).

In the remainder of this document, we describe a computer experiment designed to mimic major sources of non-sampling variability in population counts from the 2010 census. Our simulations use block-level data from *Summary File 1* (SF1) for fifty states plus the District of Columbia, combined with results from the 2010 CCM program, all of which are available to the public. Although these results are most applicable to 2010, they may also lend insight into the properties of counts from 2020 to be released later this year. Extrapolations to 2020, however, are subject to these caveats: first and foremost, the major disruptions to census operations caused by corona virus pandemic; second, the impact of new data-collection methodologies introduced in 2020, including the online self-response option and the enhanced use of administrative records and third-party data; and third, the impact of random noise added by the new Disclosure Avoidance System (DAS), which differs from the random data-swapping methods used for confidentiality protection in 2010.

In designing this simulation, we identified major sources of error in census counts and chose whether and how to describe each one based on subject matter knowledge and the data available to us. When in doubt, we tended to be conservative, erring on the side of reflecting too little noise rather than too much. Therefore, we interpret these results not as the most plausible educated guesses on the amount of natural variability in census data, but as lower bounds.

## 2 Components of a Census Population Count

### 2.1 Housing Units versus Group Quarters

For this simulation, we focus on census blocks, the smallest geographic regions for which census figures are tabulated. For the 2010 census, the United States was divided into more than 11 million blocks, many of which were uninhabited. For every census block  $b = 1, \dots, B$ , the SF1 provides

$$P_b = \text{census population count in block } b.$$

Note that  $P_b$  is not necessarily the true number of persons actually residing in the block on Census Day. The goal of this exercise is neither to estimate the true population nor to assess the likely size of the discrepancy between  $P_b$  and the true population. Rather, we are describing the random variability we would see in  $P_b$  if the census could be repeated under similar conditions with the true population held fixed.

For census purposes, the dwellings where people reside are of two fundamentally different types.

- **Group quarters (GQs).** Approximately 2.6% of the 2010 census population total came from GQs. Roughly speaking, GQs are places owned or managed by service providers where people live in a group arrangement, and residents of a GQ are typically not related to one another. Examples include prisons, dormitories, military barracks, residential treatment facilities, convents, and group homes for persons with disabilities.
- **Housing units (HUs).** The remaining 97.4% of persons in the 2010 census were counted in HUs, which include houses, apartments, and mobile homes that are occupied as someone's usual place of residence.

Census enumeration procedures for GQs and HUs are quite different, and SF1 tabulates them separately within each block. Thus we have

$$P_b = P_b^{(\text{GQ})} + P_b^{(\text{HU})},$$

where

$$\begin{aligned} P_b^{(\text{GQ})} &= \text{persons in block } b \text{ within GQs, and} \\ P_b^{(\text{HU})} &= \text{persons in block } b \text{ within HUs.} \end{aligned}$$

SF1 also provides

$$H_b = \text{number of housing units in block } b.$$

Blocks with no HUs cannot have HU persons,

$$H_b = 0 \Rightarrow P_b^{(\text{HU})} = 0,$$

but blocks with no HU persons may have HUs,

$$P_b^{(\text{HU})} = 0, \nRightarrow H_b = 0$$

because it is possible for all HUs in a block to be unoccupied.

GQs were omitted from the 2010 CCM program, and little is known about enumeration error for GQs relative to HUs. For this experiment, we do not attempt to quantify randomness in the GQ population counts. That is, we hold  $P_b^{(\text{GQ})}$  fixed and simulate variability in  $P_b^{(\text{HU})}$  for  $b = 1, \dots, B$ .

## 2.2 Substitutions

Approximately 1.9% of the HU persons in the 2010 census were substituted, meaning that their full set of characteristics (age, sex, relationship to Person 1 on the census form, race, and Hispanic origin) had been imputed or filled in. Substitution was used when the number of persons living in the HU could not be determined. In that case, the number of persons was predicted by a set of procedures known as count imputation. Substitution was also necessary when the number of persons living in the unit had been determined in some fashion (e.g., by interviewing a neighbor) but little or no additional information was available. In both of these situations, all characteristics for the persons in the unit were filled in with values borrowed from a nearby unit with the same number of persons. SF1 groups these two types together, reporting

$$I_b = \text{total number of substitutions in block } b.$$

Substitutions are traditionally excluded from census coverage studies because, even when they represent actual persons, so little is known about them that it would be difficult or impossible to determine whether and where they should have been counted. Nevertheless, the presence of a large number of substituted persons in a block suggests greater degree of noise in the HU population count, and the realized value of  $I_b$  could easily change if the census were repeated.

## 2.3 Correct and Erroneous Enumerations

For this study, we classify the non-substituted HU persons into five different types,

$$P_b^{(\text{HU})} - I_b = E_{1b} + E_{2b} + E_{3b} + E_{4b} + E_{5b},$$

where

Table 1: Estimated components of coverage for the United States housing-unit (HU) population (in thousands) in the 2010 census, based on results from the Census Coverage Measurement program. Source: Mule (2012), Table 3

|  | Estimate | Percent |
|--|----------|---------|
| Total HU population                        | 300,703  | 100.0   |
| Correct enumerations                       | 284,668  | 94.7    |
| Enumerated in correct block                | 280,852  | 93.4    |
| Mislocated to another block in same county | 2,039    | 0.7     |
| Mislocated to another county in same state | 830      | 0.3     |
| Mislocated to another state                | 948      | 0.3     |
| Erroneous enumerations                     | 10,042   | 3.3     |
| Substitutions                              | 5,993    | 2.0     |

- $E_{1b}$  represents correct enumerations (persons eligible to be counted in the census and who were indeed counted) who were correctly located in block  $b$ ;
- $E_{2b}$  represents correct enumerations who were mislocated to block  $b$  from another block in the same county;
- $E_{3b}$  represents correct enumerations who were mislocated to block  $b$  from another county in the same state;
- $E_{4b}$  represents correct enumerations who were mislocated to block  $b$  from another state; and
- $E_{5b}$  represents erroneous enumerations (fictitious persons, persons not alive or living outside the United States on Census Day, duplicates of persons who were correctly counted elsewhere).

Because  $P_b^{(\text{HU})}$  and  $I_b$  are published in SF1, the sum  $E_{1b} + \dots + E_{5b}$  is known for every block in the 2010 census, but the values of the individual components  $E_{1b}, \dots, E_{5b}$  are not. Estimates of the national totals  $\sum_{b=1}^B E_{1b}, \dots, \sum_{b=1}^B E_{5b}$  obtained from the 2010 CCM program were reported by Mule (2012) and are shown in Table 1.

### 3 Simulation Procedures

#### 3.1 Overview of Simulation

Using the decomposition

$$P_b^{(\text{HU})} = I_b + E_{1b} + E_{2b} + E_{3b} + E_{4b} + E_{5b},$$

we simulated new values of  $P_b^{(\text{HU})}$  for every block  $b$  by drawing plausible new values of the components on the right-hand side of this equation and adding them up. The simulated versions are denoted by

$$P_b^{(\text{HU})*} = I_b^* + E_{1b}^* + E_{2b}^* + E_{3b}^* + E_{4b}^* + E_{5b}^*. \quad (1)$$

For each block, the number of HUs was held fixed at  $H_b$ , the value reported in SF1. That is, we did not explicitly model noise due to omissions, erroneous inclusions, and mislocations of entire HUs that arise when constructing the MAF. However, some of that noise in the HU counts did show up in the variability of the components of the person counts in (1). For example, our simulated values of  $E_{2b}^*$ , which we describe below, were meant to include people who were mislocated to a neighboring block because the HU where they live was mistakenly placed in the wrong block on the MAF.

With  $H_b$  fixed, any block with  $H_b = 0$  must produce  $P_b^{(\text{HU})*} = 0$ . We eliminated those blocks from our universe, considering only the 6,379,963 blocks from the 2010 census that contained at least one HU, with 300,758,215 total HU persons.

#### 3.2 Simulating the Number of Substitutions

Counts of substituted persons were simulated by an empirical Bayes procedure that smoothes the estimated substitution rates across blocks within states. Suppose that

$$\begin{aligned} I_b | H_b, \lambda_b &\sim \text{Poisson}(H_b \lambda_b), \\ \lambda_b | \alpha, \beta &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

where  $\text{Poisson}(\mu)$  denotes a Poisson distribution with mean  $\mu$ , and  $\text{Gamma}(\alpha, \beta)$  denotes a Gamma distribution with shape parameter  $\alpha$  and rate parameter  $\beta$ . This is the standard Bayesian Poisson model parameterized in terms of a known exposure  $H_b$  and an unknown per-HU substitution rate  $\lambda_b$  (see Section 2.7 by Gelman et al. (2013)). For fixed values of the hyperparameters  $\alpha$  and  $\beta$ , the implied posterior distribution for the substitution rate is

$$\lambda_b | I_b, H_b, \alpha, \beta \sim \text{Gamma}(\alpha + I_b, \beta + H_b). \quad (2)$$

To obtain plausible values for  $\alpha$  and  $\beta$ , we estimated them independently within each state by the method of maximum likelihood (ML). Marginalizing over the unknown  $\lambda_b$ , the conditional distribution of  $I_b$  given  $H_b$  is negative binomial with log-mean

$$\log E(I_b | H_b) = \log H_b + \beta_0$$

and concentration parameter  $\alpha$ . Using negative binomial regression as implemented in the R package **MASS** (Venables and Ripley, 2013), we fit an intercept-only negative binomial regression with a log link and an offset term  $\log H_b$ . Fitting this model to the blocks within a state yields two estimated parameters: the intercept  $\hat{\beta}_0$  and the concentration parameter called  $\hat{\theta}$  which corresponds to  $\alpha$ . After fitting this model, we set

$$\begin{aligned}\alpha &= \hat{\theta}, \\ \beta &= \hat{\theta} / \exp(\hat{\beta}_0).\end{aligned}$$

After obtaining  $\alpha$  and  $\beta$  for a given state, we generated new versions of the block-level substitution counts by drawing

$$\lambda_b | I_b, H_b \sim \text{Gamma}(\alpha + I_b, \beta + H_b),$$

followed by

$$I_b^* | H_b, \lambda_b \sim \text{Poisson}(H_b \lambda_b)$$

independently for each block in the state. Under this procedure, the expected number of substitutions generated for any given block is a compromise between (a) the number of substitutions actually seen for that block in the 2010 census, and (b) the average number of substitutions seen for all blocks with a similar number of HUs.

Because  $\alpha$  and  $\beta$  depend only on data from SF1, they remained fixed over repetitions of the simulation, but new values for  $\lambda_b$  and  $I_b^*$  were generated for each block in each simulation run.

### 3.3 Simulating Correct and Erroneous Enumerations

Before generating  $E_{1b}^*, \dots, E_{5b}^*$ , we first needed to guess the values of  $E_{1b}, \dots, E_{5b}$  that were realized in the 2010 census. As previously mentioned, SF1 provides the total  $P_b^{(\text{HU})} - I_b = E_{1b} + \dots + E_{5b}$  for every block, and results from the CCM program provide some guidance on how to subdivide those totals. We regard the vector  $\mathbf{E}_b = (E_{1b}, E_{2b}, E_{3b}, E_{4b}, E_{5b})$  as having a multinomial distribution

$$\mathbf{E}_b | P_b^{(\text{HU})}, I_b, \boldsymbol{\pi}_b \sim \text{Mult}(P_b^{(\text{HU})} - I_b, \boldsymbol{\pi}_b), \quad (3)$$

Table 2: Mean and percentiles of random probabilities drawn from the Dirichlet distribution with concentration  $\kappa = 200$ , representing block-level heterogeneity in rates of correct and erroneous enumerations

|            | mean  | 0.5%  | 2.5%  | 25%   | 50%   | 75%   | 97.5% | 99.5% |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|
| $\pi_{1b}$ | 0.953 | 0.906 | 0.920 | 0.944 | 0.955 | 0.964 | 0.978 | 0.983 |
| $\pi_{2b}$ | 0.007 | 0.000 | 0.000 | 0.003 | 0.005 | 0.010 | 0.022 | 0.031 |
| $\pi_{3b}$ | 0.003 | 0.000 | 0.000 | 0.000 | 0.002 | 0.004 | 0.014 | 0.021 |
| $\pi_{4b}$ | 0.003 | 0.000 | 0.000 | 0.000 | 0.002 | 0.004 | 0.014 | 0.021 |
| $\pi_{5b}$ | 0.034 | 0.010 | 0.014 | 0.025 | 0.032 | 0.042 | 0.063 | 0.075 |

where  $\boldsymbol{\pi}_b = (\pi_{1b}, \pi_{2b}, \pi_{3b}, \pi_{4b}, \pi_{5b})$  is a vector of probabilities that sum to one. Based on the percentages shown in Table 1, it is reasonable to believe that

$$\begin{aligned}
\pi_{1b} &\approx 0.934/(1 - 0.02) = 0.953, \\
\pi_{2b} &\approx 0.007/(1 - 0.02) = 0.007, \\
\pi_{3b} &\approx 0.003/(1 - 0.02) = 0.003, \\
\pi_{4b} &\approx 0.003/(1 - 0.02) = 0.003, \\
\pi_{5b} &\approx 0.033/(1 - 0.02) = 0.034.
\end{aligned}$$

However, it is not reasonable to force  $\boldsymbol{\pi}_b$  to be identical for  $b = 1, \dots, B$ , as variation in local conditions will produce some block-to-block heterogeneity. Moreover, even for a single block, it seems apparent that  $\boldsymbol{\pi}_b$  could change over repeated realizations of the census, because certain local conditions (e.g., the particular enumerator assigned to the block) could vary over those realizations. To reflect this variation, we simulate  $\boldsymbol{\pi}_b$  for each block at each simulation run by drawing it from a Dirichlet distribution

$$\boldsymbol{\pi}_b \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \quad (4)$$

where  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$  is the vector of Dirichlet shape parameters. To obtain  $(0.953, 0.007, 0.003, 0.003, 0.034)$  as the average value for  $\boldsymbol{\pi}_b$ , we set the vector of shape parameters to

$$\boldsymbol{\alpha} = \kappa \times (0.953, 0.007, 0.003, 0.003, 0.034),$$

where  $\kappa > 0$  is a concentration parameter that controls how tightly the  $\boldsymbol{\pi}_b$  vectors are distributed around their average. For this experiment, we set  $\kappa = 200$ , which keeps the  $\boldsymbol{\pi}_b$  within a plausible range. Percentiles for the probabilities  $\pi_{1b}, \dots, \pi_{5b}$  under this joint distribution are shown in Table 2. For example, the rate of erroneous enumerations  $\pi_{5b}$  has a mean of 0.034, and 99% of these rates lie within  $(0.010, 0.075)$ .

In each simulation run, for each block  $b = 1, \dots, B$ , we guess the value of  $\mathbf{E}_b = (E_{1b}, E_{2b}, E_{3b}, E_{4b}, E_{5b})$  by first drawing  $\boldsymbol{\pi}_b$  from the Dirichlet distribution (4), and then drawing  $\mathbf{E}_b$  from the multinomial distribution (3). After obtaining this guess for  $\mathbf{E}_b$ , we proceed to simulate a new version that could plausibly have arisen if the census were repeated on the same population under similar conditions. The new version  $\mathbf{E}_b^* = (E_{1b}^*, E_{2b}^*, E_{3b}^*, E_{4b}^*, E_{5b}^*)$  was created as follows.

1. Set  $E_{1b}^* = E_{1b}$ .<sup>1</sup>
2. Distribute  $E_{2b}$  persons over randomly selected blocks in the same county, drawing these blocks with probability proportional to size with replacement (ppswr), using the number of housing units in the block  $H_b$  as the measure of size. After performing this distribution for all blocks in the county, set  $E_{2b}^*$  to the number placed into block  $b$ .<sup>2</sup>
3. Distribute  $E_{3b}$  persons over blocks in the same state, selecting the blocks by ppswr. After performing this distribution for all blocks in the state, set  $E_{3b}^*$  to the number placed into block  $b$ .
4. Distribute  $E_{4b}$  persons over blocks in the nation, selecting the blocks by ppswr. After performing this distribution for all blocks in the nation, set  $E_{4b}^*$  to the number placed into block  $b$ .
5. Simulate  $E_{5b}^*$  by drawing from a Poisson distribution with mean  $P_b^{(\text{HU})} \times \pi_{5b}$ .<sup>3</sup> Repeat for blocks  $b = 1, \dots, B$ .

Note that in Steps 2–4, we are using  $E_{2b}$ ,  $E_{3b}$  and  $E_{4b}$ , the guessed numbers of persons mislocated *into* block  $b$  from other places, as proxies for the unknown numbers of persons mislocated *out of* block  $b$  to other places. In effect, we remove  $E_{2b} + E_{3b} + E_{4b}$  persons from the block and replace them by  $E_{2b}^* + E_{3b}^* + E_{4b}^*$  persons. The

<sup>1</sup>Recall that  $E_{1b}$  represents persons who were captured by the census in the correct block. The easy-to-count population is concentrated in this group, and in a repetition of the census, it is likely that many of those persons would again be captured in the correct block. However,  $E_{ib}$  may also include some hard-to-count persons who happened to be captured in the actual census but might not be captured in a repetition. Therefore, setting  $E_{1b}^* = E_{1b}$  is conservative, in the sense that it will tend to understate the variability in  $E_{1b}^*$ , particularly in areas that have many persons who are prone to be missed, but whose probabilities of being counted are not so low that they could never be captured.

<sup>2</sup>This procedure, which mislocates each person independently, ignores the fact that some mislocations happen at the HU level, when an HU is mistakenly placed in an adjacent block on the MAF, causing all persons in the HU to be mislocated together. A simulation involving HU-level mislocations would tend to increase the variance of  $E_{2b}^*$ , so our procedure is conservative.

<sup>3</sup>The marginal distribution of  $E_{5b}$  implied by (3) is binomial. Drawing  $E_{5b}$  from that binomial would force the simulated value  $E_{5b}^*$  to be less than or equal to the number of enumerated HU persons reported in the 2010 census, which is logically unnecessary. The Poisson distribution used for  $E_{5b}^*$  has the same mean as the binomial, a slightly higher variance, and no strict upper limit.

number removed has extra-binomial variability arising from the Dirichlet sampling of  $\pi_b$ , whereas the number added is essentially a Poisson variate with mean proportional to  $H_b$ .

This simulation was programmed in R and run 100 times using R version 3.5.2 (64-bit) on Redhat linux.

## 4 Results

### 4.1 Measures of Error

Consider a geographic domain  $D$  that corresponds to a set of blocks. Let

$$\mathcal{N}_D = \sum_{b \in D} P_b^{(\text{HU})}$$

denote the published 2010 census count of HU persons, and let

$$\mathcal{N}_D^* = \sum_{b \in D} P_b^{(\text{HU})^*}$$

denote the count of HU persons generated by a simulation run. The *mean absolute deviation* (MAD) is

$$E(|\mathcal{N}_D^* - \mathcal{N}_D|), \quad (5)$$

where the expectation is taken over repeated simulation runs. The *mean absolute percent error* (MAPE) is defined as

$$100 \times E\left(\frac{|\mathcal{N}_D^* - \mathcal{N}_D|}{\mathcal{N}_D}\right), \quad (6)$$

which is defined if  $\mathcal{N}_D > 0$ . To estimate MAD and MAPE, we replace the expectations in (5) and (6) by sample averages over the 100 runs.

To summarize over a collection of domains, we average the estimated MAD and MAPE over that collection. We also examine the raw signed error (RSE) ( $\mathcal{N}_D^* - \mathcal{N}_D$ ), computing percentiles of the RSE over the collection of domains within each simulation run, then averaging those percentiles over the runs. Percentiles of RSE will reveal situations where the error distributions are asymmetric, which tends to happen for small domains (e.g., blocks) where  $\mathcal{N}_D$  may be nearly or exactly zero.

Error summaries for the entire United States, each of the fifty states, and the District of Columbia are shown in Table 3. At the national level, the typical error (MAD) is approximately  $\pm 4,716$  persons, or 0.002% of the HU population. At the state

Table 3: Error summaries for the United States, fifty states and the District of Columbia: housing unit population from the 2010 census, mean absolute deviation (MAD), and mean absolute percent error (MAPE) over 100 simulation runs

|                      | Population  | MAD    | MAPE  |
|----------------------|-------------|--------|-------|
| United States        | 300,758,215 | 4,716  | 0.002 |
| Alabama              | 4,663,920   | 1,006  | 0.022 |
| Alaska               | 683,879     | 204    | 0.030 |
| Arizona              | 6,252,633   | 1,130  | 0.018 |
| Arkansas             | 2,836,987   | 598    | 0.021 |
| California           | 36,434,140  | 15,642 | 0.043 |
| Colorado             | 4,913,318   | 732    | 0.015 |
| Connecticut          | 3,455,945   | 566    | 0.016 |
| Delaware             | 873,521     | 258    | 0.030 |
| District of Columbia | 561,702     | 402    | 0.072 |
| Florida              | 18,379,601  | 6,853  | 0.037 |
| Georgia              | 9,434,454   | 933    | 0.010 |
| Hawaii               | 1,317,421   | 413    | 0.031 |
| Idaho                | 1,538,631   | 405    | 0.026 |
| Illinois             | 12,528,859  | 1,566  | 0.012 |
| Indiana              | 6,296,879   | 667    | 0.011 |
| Iowa                 | 2,948,243   | 433    | 0.015 |
| Kansas               | 2,774,044   | 360    | 0.013 |
| Kentucky             | 4,213,497   | 607    | 0.014 |
| Louisiana            | 4,405,945   | 665    | 0.015 |
| Maine                | 1,292,816   | 996    | 0.077 |
| Maryland             | 5,635,177   | 723    | 0.013 |
| Massachusetts        | 6,308,747   | 679    | 0.011 |
| Michigan             | 9,654,572   | 2,026  | 0.021 |
| Minnesota            | 5,168,530   | 760    | 0.015 |
| Mississippi          | 2,875,333   | 452    | 0.016 |
| Missouri             | 5,814,785   | 1,085  | 0.019 |
| Montana              | 960,566     | 498    | 0.052 |
| Nebraska             | 1,775,176   | 312    | 0.018 |
| Nevada               | 2,664,397   | 529    | 0.020 |
| New Hampshire        | 1,276,366   | 413    | 0.032 |
| New Jersey           | 8,605,018   | 1,751  | 0.020 |
| New Mexico           | 2,016,550   | 446    | 0.022 |
| New York             | 18,792,424  | 1,217  | 0.006 |
| North Carolina       | 9,278,237   | 2,101  | 0.023 |
| North Dakota         | 647,535     | 213    | 0.033 |
| Ohio                 | 11,230,238  | 1,417  | 0.013 |
| Oklahoma             | 3,639,334   | 593    | 0.016 |
| Oregon               | 3,744,432   | 527    | 0.014 |
| Pennsylvania         | 12,276,266  | 1,163  | 0.009 |
| Rhode Island         | 1,009,904   | 275    | 0.027 |
| South Carolina       | 4,486,210   | 1,137  | 0.025 |
| South Dakota         | 780,130     | 255    | 0.033 |
| Tennessee            | 6,192,633   | 910    | 0.015 |
| Texas                | 24,564,422  | 4,608  | 0.019 |
| Utah                 | 2,717,733   | 1,475  | 0.054 |
| Vermont              | 600,412     | 390    | 0.065 |
| Virginia             | 7,761,190   | 662    | 0.009 |
| Washington           | 6,585,165   | 646    | 0.010 |
| West Virginia        | 1,803,612   | 588    | 0.033 |
| Wisconsin            | 5,536,772   | 1,279  | 0.023 |
| Wyoming              | 549,914     | 260    | 0.047 |

Table 4: Error summaries for United States counties, classified by the population of housing unit (HU) persons: number of counties (N), mean absolute deviation (MAD), mean absolute percent error (MAPE), and percentiles of raw signed error (RSE), averaged over 100 simulation runs

|                                       | N     | MAD   | MAPE | RSE    |        |        |      |     |       |       |
|---------------------------------------|-------|-------|------|--------|--------|--------|------|-----|-------|-------|
|                                       |       |       |      | 0.5%   | 2.5%   | 25%    | 50%  | 75% | 97.5% | 99.5% |
| All counties                          | 3,143 | 117   | 0.31 | -1,603 | -476   | -29    | 17   | 67  | 356   | 866   |
| Counties with HU pop. < 1,000         | 37    | 10    | 1.60 | -16    | -12    | -1     | 5    | 12  | 31    | 36    |
| Counties with HU pop. 1,000–9,999     | 691   | 28    | 0.56 | -78    | -51    | -5     | 14   | 34  | 86    | 120   |
| Counties with HU pop. 10,000–99,999   | 1,849 | 74    | 0.26 | -307   | -180   | -28    | 24   | 76  | 221   | 340   |
| Counties with HU pop. 100,000–999,999 | 527   | 292   | 0.11 | -1,723 | -1,147 | -211   | -14  | 172 | 726   | 1,250 |
| Counties with HU pop. 1,000,000+      | 39    | 1,463 | 0.08 | -5,584 | -4,664 | -1,833 | -631 | 583 | 1,660 | 2,381 |

level, the MAD ranges from  $\pm 204$  persons (Alaska) to  $\pm 15,642$  persons (California), and the MAPE ranges from 0.006% (New York) to 0.077% (Maine).

Table 4 shows error summaries for counties, grouping them by population of HU persons using cutpoints that are equally spaced on a log scale. In addition to MAD and MAPE, this table also displays select percentiles of the RSE. Averaged over all counties, the typical error (MAD) is  $\pm 117$  persons, or 0.31% of the population. MAD tends to increase with county population size, whereas MAPE tends to decrease. Percentiles of the RSE reveal an interesting pattern: the distribution of raw signed errors is right-skewed for smaller counties and left-skewed for larger counties.

Table 5 shows summary measures at the block level. Blocks are grouped by the number of HU persons in the 2010 census. MAPE is omitted from this table, because the statistic is undefined for blocks with zero HU persons. Averaged over all blocks, the typical error is  $\pm 1.5$  persons. In blocks with zero HU persons, the distribution of RSE is right-skewed, because simulated population counts cannot be negative. In all other block size categories, the RSE is nearly symmetric.

To reiterate, we believe that the summaries in these tables should be viewed as lower bounds on the natural variability in the census counts. Some sources of variation were not represented in the simulation, notably:

- Group quarters were not considered, eliminating noise in all GQ population counts.
- The number of HUs in each block was held constant, eliminating some of the noise due to errors in the listing of HUs.
- Setting  $E_{1b}^* = E_{1b}$  very likely understated the amount of noise in these counts of correctly enumerated persons, because randomness in the number of omissions was not considered.

Table 5: Error summaries for United States census blocks containing at least one housing unit (HU), classified by the population of HU persons: number of blocks (N), mean absolute deviation (MAD), and percentiles of raw signed error (RSE), averaged over 100 simulation runs

|                                 | N         | MAD  | RSE  |      |     |     |     |       |       |
|---------------------------------|-----------|------|------|------|-----|-----|-----|-------|-------|
|                                 |           |      | 0.5% | 2.5% | 25% | 50% | 75% | 97.5% | 99.5% |
| All blocks with at least one HU | 6,379,963 | 1.5  | -10  | -5   | -1  | 0   | 1   | 5     | 10    |
| Blocks with 0 HU persons        | 191,885   | 0.1  | 0    | 0    | 0   | 0   | 0   | 1     | 2     |
| Blocks with 1–9 HU persons      | 1,825,130 | 0.4  | -3   | -2   | 0   | 0   | 0   | 2     | 3     |
| Blocks with 10–99 HU persons    | 3,663,268 | 1.5  | -7   | -4   | -1  | 0   | 1   | 4     | 7     |
| Blocks with 100–999 HU persons  | 691,557   | 4.1  | -18  | -11  | -3  | 0   | 3   | 12    | 19    |
| Blocks with 1,000+ HU persons   | 8,123     | 14.5 | -57  | -39  | -11 | 1   | 12  | 37    | 53    |

- Mislocated individuals in each block were independently distributed across other blocks, ignoring the clustering that occurs when all residents of a unit are misplaced together.

## References

- Fellegi, I. P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, 59(308):1016–1041.
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Press, Boca Raton, FL.
- Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38(2):359–374.
- Mule, T. (2012). Census coverage measurement estimation report: Summary of estimates of coverage for persons in the United States. Technical Report DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01, Department of Commerce, U.S. Census Bureau, Washington, DC.
- Tepping, B. J. and Bailar, B. A. (1973). Enumerator variance in the 1970 census. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 160–169.
- Venables, W. N. and Ripley, B. D. (2013). *Modern Applied Statistics with S, Fourth Edition*. Springer Science & Business Media, New York.

## Attachment 2: Simulating Block Level Populations Using 2010 Census Data and Coverage Measurement Results

William Bell  
April 19, 2021

The formula for dual system estimation (DSE) of the housing unit population<sup>1</sup> is

$$\begin{aligned} \text{DSE} &= \text{Cen} \times r_{DD} \times r_{CE}/r_M \\ &= \text{Cen} \times \frac{DD}{\text{Cen}} \times \frac{CE}{DD} \bigg/ \frac{M}{P} \end{aligned} \tag{1}$$

where  $\text{Cen}$  = census count,  $DD$  = number of data defined census enumerations,  $CE$  = number of correct enumerations,  $P$  = P-sample total, and  $M$  = P-sample matches. (The P-sample is the sample of persons and housing units from the post-enumeration survey (PES) done to measure census coverage by matching results to the census.) If the independence assumption that underlies the use of this formula holds, then with a large P-sample used to estimate the match rate,  $r_M \rightarrow CE/N$ , where  $N$  is the true population. In this case  $\text{DSE} \rightarrow N$ . Here  $DD$ , and hence  $r_{DD}$ , are treated as known so that, under the independence assumption, errors in estimating  $r_{CE}$  and  $r_M$  are what produce errors in DSE as an estimate of the true housing unit population  $N$ .<sup>2</sup>

Conversely, from a Bayesian perspective, uncertainty about  $r_{CE}$  and  $r_M$  are what lead to uncertainty about  $N$ . From this perspective, simulations of  $r_{CE}$  and  $r_M$  that reflect uncertainty about these quantities can be used with formula (1) to provide simulations of the true population  $N$  that reflect the uncertainty about  $N$ . If the simulations came from posterior distributions obtained from modeling these rates given suitable data, these would be posterior simulations of  $N$  given the data. With simulations obtained just using distributional assumptions about  $r_{CE}$  and  $r_M$ , these may be better termed “prior simulations.”

The simulations of  $r_M$  create heterogeneity that, in the simulations described here, led to overall simulated positive undercounts. This is analogous to unaccounted for variation in match rates estimated in the PES leading to “correlation bias” in the DSEs (for which the PES uses demographic analysis sex ratios to make an adjustment to the DSEs for adult males). Because the simulations reported here don’t account

---

<sup>1</sup>In 2000 and 2010, DSE was not applied to the group quarters population.

<sup>2</sup>Whole person imputations (non-data defined person records), which are  $\text{Cen} - DD$ , are effectively estimates for persons in housing units for which no, or very limited, information was supplied about the residents. They are subject to their own errors as estimates for the actual residents, but as these records do not contain enough information for the PES matching and follow-up, they are subtracted from  $\text{Cen}$  and effectively treated as (potential) census omissions by DSE.

for known heterogeneity across demographic groups and other factors, the simulated populations are rescaled to force agreement with the census total at the national level. This is consistent with the estimated net undercount in the 2010 census being very close to zero.

The DSE formula (1) was applied using block level simulations of  $r_{CE}$  and  $r_M$  to produce simulations of block level populations via the following steps.

1. For each block  $b$ , start with the actual census count of data defined persons,  $DD_b$ .
2. Simulate correct enumerations for the block,  $CE_b$ , by first simulating the block's correct enumeration probability,  $r_{CE,b}$ , then simulating the block's count of correct enumerations from a Binomial( $DD_b, r_{CE,b}$ ) distribution. Details of these simulations are given in Joe Schafer's documentation (Attachment 1) of how he simulated various components of the census count.<sup>3</sup> For the purposes here, imputations are not simulated; we hold the census count and its number of data defined enumerations fixed.
3. Simulate the block's census inclusion probability,  $r_{M,b}$ , from a Beta( $\alpha, \beta$ ) distribution with  $\alpha = 6$  and  $\beta = .7$ . These values of  $\alpha$  and  $\beta$  were chosen to provide what we believed to be a reasonable amount of variation over blocks in the census inclusion probabilities, while giving a median of .933, this figure closely matching the overall probability of census inclusion *in the correct block* obtained from the 2010 PES.<sup>4</sup>
4. Use the block simulated values of  $CE_b$  and  $r_{M,b}$  to calculate an initial simulated value of block population from the DSE formula as  $Pop1_b = CE_b / r_{M,b}$ . Rescale all the block level  $Pop1_b$  so their sum across all blocks agrees with the national census count of the housing unit population, i.e., compute  $Pop2_b = Pop1_b \times (\sum_j Cen_j / \sum_j Pop1_j)$ .
5. Round  $Pop2_b$  to an integer; call this  $Pop3_b$ . Simulate census omissions for block  $b$  from a Binomial( $Pop3_b, 1 - r_{M,b}$ ) distribution. Add these to the correct

---

<sup>3</sup>Note that we define CE as enumerated in the correct block (actually, within the block cluster's search area), and do not need the components of census enumerations that Joe Schafer simulated that refer to counted in the right county but wrong block, etc.

<sup>4</sup>Table 3 of DSSD 2010 Census Coverage Measurement Memorandum #2010-G-01 (Mule 2012) reports an overall omission rate of .053 corresponding to a probability of census inclusion of .947. However, this is the probability of someone being included anywhere in the U.S. For the simulation of block level population we want the probability of someone being included in the census in the correct block. From Table 3 this is  $280,852/300,667 = .934$ , where 280,852,000 is the estimated number of census enumerations in the correct blocks, and 300,667,000 is the total estimate of the housing unit population from DSE.

census enumerations within the block to get revised block level housing unit populations, i.e.,  $\text{Pop4}_b = CE_b + \text{omissions}_b$ .

6. As in Step 4, rescale the  $\text{Pop4}_b$  values to get revised population figures,  $\text{Pop5}_b = \text{Pop4}_b \times (\sum_j \text{Cen}_j / \sum_j \text{Pop4}_j)$ .
7. Round the  $\text{Pop5}_b$  to produce  $\text{Pop}_b$ , the final simulated block level populations. Aggregate these to counties, states, and the nation.

The simulation of the block level correct enumeration and census inclusion probabilities,  $r_{CE,b}$  and  $r_{M,b}$ , at Steps 2 and 3 provides plausible variation of these quantities over blocks. Simulation of the counts of correct enumerations and census omissions,  $CE_b$  and  $\text{omissions}_b$ , at Steps 2 and 5, then provides additional variation in the realized block level  $CE$  and omission proportions. These binomial simulations sensibly provide more relative variation in the results for smaller blocks, since the distribution of a proportion from a binomial distribution becomes more concentrated the larger is the binomial  $n$  (which here is either  $DD_b$  or  $\text{Pop3}_b$ ).

The rescaling of the simulated block population values at Steps 4 and 6 so they sum to the national census count was done to control the net national undercount in the simulated populations to be close to zero, as was the case for the 2010 census.<sup>5</sup> When the parameters of the beta distribution used to simulate the  $r_{M,b}$  were determined so the beta distribution mean, rather than its median, matched the overall P-sample match rate, more substantial undercounts resulted, so that the rescaling of the simulated population values at Step 4 was more severe. This alternative was thus seen as less desirable.

The other alternative of simply analyzing simulation results that reflect an overall national undercount of some substance, when this was not observed for the 2010 census, was also seen as less desirable. Previous research has established how  $CE$  rates and P-sample match rates vary with certain population characteristics, such as demographics, and none of that variation is reflected in the simple simulation framework used here. What can be learned from the simulations is something about how coverage error may vary across subnational areas due to variations in key quantities – correct enumeration rates and census inclusion probabilities. The simulations provide no new information about the overall level of undercount or overcount, though the choices of how much variation is allowed in the key quantities do affect this in the simulations. Thus, it seemed sensible to control the net national undercount to remain close to a value near zero as was the case for the 2010 census and PES from which the data used here was drawn.

---

<sup>5</sup>The net national undercount if we compare census counts to the  $\text{Pop2}_b$  from Step 4 is zero by construction. The additional variation from the binomial simulations of omissions at Step 5 then produces a national undercount that motivates the rescaling at Step 6. After rounding the  $\text{Pop5}_b$  values at Step 7, the final  $\text{Pop}_b$  values yield a small national level undercount of about 0.08%.

Limitations of the simulation results include the following:

- Variation in the block level  $r_{CE}$  and  $r_M$  rates is obtained by simulations from assumed distributions. These distributions do not reflect known heterogeneity in the rates across demographic groups or other factors. The assumed distributions are just an attempt to reflect a reasonable amount of variation in the block level rates as an aggregate. Note that the Appendix shows that distributions across blocks of the simulated  $CE/DD$  and  $CE/N$  proportions agree well with tabulations of the distributions of block level CE and P-sample match rates from the 2010 PES. The simulations do not, however, provide meaningful results for specific individual states, counties, or blocks.
- For blocks with a census count of zero, or whose census count is all imputations, the DSE formula produces a population estimate of zero regardless of the values of the CE and match rates.
- The simulations of the rates do not reflect any dependence between the CE and P-sample match rates, nor possible dependence between the match rates and the block level imputation rates. There has been very little study of this subject, and so we have essentially no guidance for building any dependence into the simulations.
- Imputations are known and are effectively treated as omissions, consistent with how the DSE is applied and was used to estimate omissions for components of error in the 2010 census.<sup>6</sup> Some imputations are certainly errors in the census count (e.g., person records imputed into housing units that were vacant on census day), while others may reflect actual persons for whom very little information was reported (e.g, an accurate population count is obtained for a housing unit but with no information reported on the specific residents). Though we do not simulate imputation errors, they are to some extent covered by the DSE formula treating them as omissions.
- The DSE formula (1) is being used to simulate variation only in the housing unit population, omitting from consideration the population residing in group quarters. This is consistent with the restriction to the housing unit population of coverage measurement for the 2000 and 2010 censuses.

---

<sup>6</sup>See again Table 3 of Mule (2012), DSSD 2010 Census Coverage Measurement Memorandum #2010-G-01.

## Results from the simulations

The simulation procedure described above was applied to examine variation in coverage errors and relative percent coverage errors defined, respectively, as

$$\begin{aligned}\text{census coverage error} &= \text{Pop} - \text{Cen} \\ \text{relative coverage error} &= 100 \times \frac{\text{Pop} - \text{Cen}}{\text{Cen}}\end{aligned}$$

where, as mentioned above, both Pop and Cen refer to the housing unit population, i.e., excluding the group quarters population. The relative coverage error is much like the usual undercount rate, except it is expressed relative to the census count in the denominator, rather than to the population figure, which would give the standard definition of undercount rate. For the rough assessments made here this distinction makes little difference, and it had the advantage that the same set of denominators was used for each simulation, since the census counts are held fixed across simulations.

The simulations were carried out at the block level, with the simulated Pop values then aggregated to counties, states, and the total U.S. As noted earlier, for blocks with a census count of zero, or whose census count is all imputations, the DSE formula produces a simulated population figure of zero for every simulation, so that the coverage error for all such blocks is automatically zero. Relative coverage errors are thus undefined for such blocks and, partly for this reason, we shall not examine relative coverage errors at the block level.<sup>7</sup>

We report here summary measures of coverage error at the county and block level. The summary measures – mean absolute error, mean absolute percentage error (county level only), and a set of percentiles of the distribution of coverage error – are analogous to those used to measure variation in census counts for the simulations described in Attachment 1, which gives formal definitions of the measures. The simulations described there investigate potential variability in census counts, not census coverage errors, but while the definition of individual errors differ, the summary measures applied to the two sets of errors are the same. Because the magnitude of the coverage error is larger for larger areas, in addition to presenting results that average over all U.S. counties, we present results that average across groups of counties defined by size as determined by their census housing unit populations. Similarly, results are shown for all blocks with at least one housing unit, and for groups of blocks defined by their census housing unit person counts.

---

<sup>7</sup>It is also the case that, due to the very small populations (often single digits) of very small blocks, very large relative errors can occur at the block level. For example, if a block has one census housing unit and the census records one resident while the simulated population figure is two, that is a 100% relative undercoverage error, while if the reverse occurs (the census counts two residents and the simulation gets one), that is a 50% relative overcoverage error. The many extremes that occur for small blocks make interpretation of summary measures of relative coverage errors difficult at the block level.

Table 1 shows the summary census coverage error measures at the county level. The mean absolute error (MAE) over all counties is 964 persons; the MAEs increase substantially as the size categories increase, as was just noted. The mean absolute percentage error (MAPE) across all counties is 1.56%, and the MAPEs decrease as the size categories increase. While the extreme percentiles reflect some large in magnitude errors for the largest size category, these are for counties with very large census counts, so the MAPE for this category is only 0.77%.

**Table 1.** Simulated census coverage error measures for counties

| Subsets of counties    | $N$   | MAE    | MAPE  | percentiles of errors |         |         |
|------------------------|-------|--------|-------|-----------------------|---------|---------|
|                        |       |        |       | 0.5%                  | 2.5%    | 5%      |
| All counties           | 3,143 | 964    | 1.56% | -14,316               | -4,009  | -1,841  |
| HU pop $< 1K$          | 37    | 23     | 3.47% | -35                   | -27     | -22     |
| HU pop $1K - 10K$      | 691   | 121    | 2.32% | -366                  | -210    | -146    |
| HU pop $10K - 100K$    | 1,849 | 446    | 1.40% | -2,334                | -1,208  | -832    |
| HU pop $100K - 1,000K$ | 527   | 2,930  | 1.06% | -20,682               | -10,204 | -7,222  |
| HU pop $1,000K+$       | 39    | 14,848 | 0.77% | -54,971               | -50,462 | -44,833 |

**Table 1.** (continued) Simulated census coverage error measures for counties

| Subsets of counties    | percentiles of errors |       |        |        |        |        |
|------------------------|-----------------------|-------|--------|--------|--------|--------|
|                        | 25%                   | 50%   | 75%    | 95%    | 97.5%  | 99.5%  |
| All counties           | -86                   | 126   | 428    | 2,048  | 3,731  | 9,595  |
| HU pop $< 1K$          | 0                     | 15    | 30     | 54     | 62     | 74     |
| HU pop $1K - 10K$      | 6                     | 79    | 156    | 284    | 331    | 428    |
| HU pop $10K - 100K$    | -106                  | 178   | 449    | 1,053  | 1,316  | 1,891  |
| HU pop $100K - 1,000K$ | -1,631                | 531   | 2,059  | 6,278  | 7,910  | 10,611 |
| HU pop $1,000K+$       | -12,052               | 1,209 | 10,126 | 20,007 | 24,326 | 51,602 |

HU = housing unit, MAE = Mean Absolute (L1) (coverage) Error, MAPE = Mean Absolute Percentage coverage Error

Source: Comparisons of 2010 SF1 census counts with author's simulated population figures.

Table 2 shows the simulated coverage error measures for blocks. The MAE over all blocks is around 5.5 persons, the value depending slightly on whether one defines “all” as all blocks with at least one census housing unit, or as all blocks with at least one census person record (which could be imputed) in a housing unit. As with counties, the MAEs increase substantially with the block size categories. As noted above, MAPEs can be unstable for small blocks, and so are not shown.

**Table 2.** Simulated census coverage error measures for blocks

| Subsets of blocks           | $N$       | MAE   | percentiles of errors |      |      |
|-----------------------------|-----------|-------|-----------------------|------|------|
|                             |           |       | 0.5%                  | 2.5% | 5%   |
| All with number of HU $> 1$ | 6,379,963 | 5.4   | −47                   | −18  | −11  |
| All with HU pop $> 1$       | 6,188,078 | 5.6   | −48                   | −18  | −12  |
| HU pop 1 – 9                | 1,825,130 | 0.6   | −4                    | −2   | −1   |
| HU pop 10 – 99              | 3,663,268 | 4.4   | −15                   | −10  | −8   |
| HU pop 100 – 999            | 691,557   | 23.2  | −97                   | −57  | −42  |
| HU pop 1,000+               | 8,123     | 154.7 | −430                  | −293 | −242 |

**Table 2.** (continued) Simulated census coverage error measures for blocks

| Subsets of blocks            | percentiles of errors |     |     |     |       |       |
|------------------------------|-----------------------|-----|-----|-----|-------|-------|
|                              | 25%                   | 50% | 75% | 95% | 97.5% | 99.5% |
| All with number of HUs $> 1$ | −3                    | 0   | 1   | 12  | 23    | 73    |
| All with HU pop $> 1$        | −3                    | 0   | 1   | 13  | 24    | 74    |
| HU pop 1 – 9                 | 0                     | 0   | 0   | 2   | 3     | 5     |
| HU pop 10 – 99               | −3                    | −1  | 1   | 12  | 19    | 41    |
| HU pop 100 – 999             | −16                   | −8  | 6   | 61  | 97    | 222   |
| HU pop 1,000+                | −135                  | −78 | 31  | 388 | 591   | 1,210 |

HU = housing unit, MAE = Mean Absolute (L1) (coverage) Error

Source: Comparisons of 2010 SF1 census counts with author’s simulated population figures.

Despite the small number of simulations performed (25), the simulation error at the block level was very small. In fact, the amount of variation across simulations for the percentiles was none or very close to none for all but the largest block size category due to the very large numbers of blocks in all the other block size categories. Even for the HU pop 1,000+ category the simulations of the three middle percentiles – 25%, 50%, and 75% – had very little variation over simulations, as was also the case for the MAEs. There was more variation across simulations for the county results, though still generally quite small in a relative sense. For example, the highest Monte Carlo standard error for a MAPE was about 0.04% for the group of counties with population less than 1,000. For all the other county size groups, the Monte Carlo standard error on the MAPE was less than 0.01%.

**Appendix:** Comparing the distributions of simulated block *CE* and census inclusion proportions to the distributions of block level *CE* and P-sample match rates from the 2010 PES

The 2010 PES E-sample – a sample of the census enumerations of 5,687 blocks – was matched to the PES sample drawn from the same blocks, with nonmatches followed up to determine if they were correct or erroneous census enumerations. Sample weighted estimates of the *CE/DD* proportions were obtained for these individual blocks, and used to estimate the distribution of the block level *CE/DD* proportions. A limitation of this is the fact that the E-sample weighting was designed to scale up person estimates to represent populations for larger areas, not to scale up calculations for individual blocks to represent the universe of census blocks. Nonetheless, this calculation should provide a general indication of how much block level *CE/DD* proportions can vary. Table 3 shows a number of percentiles of this distribution in comparison to percentiles of the corresponding distribution obtained from a simulation of *CEs* for all 2010 census blocks with positive data defined counts, done as described above. We see very close agreement between the two distributions, indicating that the simulations produce a realistic amount of variation of the block level *CE/DDs*.

**Table 3. Percentiles of block level CE proportions**

| Level      | 2010 PES | Simulated population |
|------------|----------|----------------------|
| 100% (max) | 1.00     | 1.00                 |
| 99%        | 1.00     | 1.00                 |
| 95%        | 1.00     | 1.00                 |
| 90%        | 1.00     | 1.00                 |
| 75%        | 1.00     | 1.00                 |
| 50%        | .97      | .97                  |
| 25%        | .93      | .93                  |
| 10%        | .87      | .89                  |
| 5%         | .83      | .84                  |
| 1%         | .64      | .67                  |
| 0% (min)   | 0.00     | 0.00                 |

Table 4 compares distribution percentiles of block level 2010 P-sample match rates,  $r_M$ , estimated using 5,664 blocks of P-sample data, with corresponding percentiles of ratios  $CE/N$  (simulated census correct enumerations over simulated true population) from one simulated population. (Under the independence assumption used for DSE,  $r_M$  estimates  $CE/N$ ). While the agreement between the two is not as striking as for the  $r_{CE} = CE/DD$  proportions in Table 3, there is still good agreement. It should also be noted that, given the large number of blocks involved in the simulations (about 6.2 million with at least one data defined census person in a housing unit), the distribution of the simulated rates does not vary much across the simulations.

**Table 4. Percentiles of block level  $r_M$  and  $CE/N$  proportions**

| Level      | 2010 PES $r_M$ | Simulated population $CE/N$ |
|------------|----------------|-----------------------------|
| 100% (max) | 1.00           | 1.00                        |
| 99%        | 1.00           | 1.00                        |
| 95%        | 1.00           | 1.00                        |
| 90%        | 1.00           | 1.00                        |
| 75%        | .97            | 1.00                        |
| 50%        | .93            | .96                         |
| 25%        | .86            | .85                         |
| 10%        | .77            | .73                         |
| 5%         | .69            | .66                         |
| 1%         | .33            | .50                         |
| 0% (min)   | 0.00           | 0.09                        |

## References

- Mule, T. (2012). Census coverage measurement estimation report: Summary of estimates of coverage for persons in the United States. DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-01. Department of Commerce, U.S. Census Bureau, Washington, DC.

## Additional Detailed Bell and Schafer Simulation Results—Counties

Population Error (raw signed error, Simulated—2010 Census enumeration—Counties).

| Counties by size  | Number of counties | Mean absolute (L1) error (people) | Mean absolute error (percent) | Error Distribution Quantiles |               |               |            |            |            |              |              |              |
|---|--------------------|-----------------------------------|-------------------------------|------------------------------|---------------|---------------|------------|------------|------------|--------------|--------------|--------------|
|   |                    |                                   |                               | q0.005                       | q0.025        | q0.050        | q0.250     | q0.500     | q0.750     | q0.950       | q0.975       |              |
| <b>Bell Simulation (Housing Units [HU] population only)</b> |                    |                                   |                               |                              |               |               |            |            |            |              |              |              |
| <b>All counties.....</b>                                    | <b>3,143</b>       | <b>964</b>                        | <b>1.56</b>                   | <b>-14,316</b>               | <b>-4,009</b> | <b>-1,841</b> | <b>-86</b> | <b>126</b> | <b>428</b> | <b>2,048</b> | <b>3,731</b> | <b>9,595</b> |
| Counties with HU population < 1,000.....                    | 37                 | 23                                | 3.47                          | -35                          | -27           | -22           | 0          | 15         | 30         | 54           | 62           | 74           |
| Counties with HU population 1,000-9,999.....                | 691                | 121                               | 2.32                          | -366                         | -210          | -146          | 6          | 79         | 156        | 284          | 331          | 428          |
| Counties with HU population 10,000-99,999.....              | 1,849              | 446                               | 1.40                          | -2,334                       | -1,208        | -832          | -106       | 178        | 449        | 1,053        | 1,316        | 1,891        |
| Counties with HU population 100,000-999,999.....            | 527                | 2,930                             | 1.06                          | -20,682                      | -10,204       | -7,222        | -1,631     | 531        | 2,059      | 6,278        | 7,910        | 10,611       |
| Counties with HU population 1,000,000+.....                 | 39                 | 14,848                            | 0.77                          | -54,971                      | -50,462       | -44,833       | -12,052    | 1,209      | 10,126     | 20,007       | 24,326       | 51,602       |
| <b>Schafer Simulation (HU population only)</b>              |                    |                                   |                               |                              |               |               |            |            |            |              |              |              |
| <b>All counties.....</b>                                    | <b>3,143</b>       | <b>117</b>                        | <b>0.31</b>                   | <b>-1,603</b>                | <b>-476</b>   | <b>-248</b>   | <b>-29</b> | <b>17</b>  | <b>67</b>  | <b>230</b>   | <b>356</b>   | <b>866</b>   |
| Counties with HU population < 1,000.....                    | 37                 | 10                                | 1.60                          | -16                          | -12           | -10           | -1         | 5          | 12         | 27           | 31           | 36           |
| Counties with HU population 1,000-9,999.....                | 691                | 28                                | 0.56                          | -78                          | -51           | -38           | -5         | 14         | 35         | 71           | 86           | 120          |
| Counties with HU population 10,000-99,999.....              | 1,849              | 74                                | 0.26                          | -307                         | -180          | -131          | -28        | 24         | 76         | 177          | 221          | 340          |
| Counties with HU population 100,000-999,999.....            | 527                | 292                               | 0.11                          | -1,723                       | -1,147        | -784          | -211       | -14        | 172        | 545          | 726          | 1,250        |
| Counties with HU population 1,000,000+.....                 | 39                 | 1,463                             | 0.08                          | -5,584                       | -4,664        | -3,659        | -1,833     | -631       | 583        | 1,351        | 1,660        | 2,381        |

Source: U.S. Census Bureau, Research and Methodology Directorate.

## Additional Detailed Bell and Schafer Simulation Results—Blocks

Population Error (raw signed error, Simulated—2010 Census enumeration—Blocks).

| Blocks by size  | Number of blocks | Mean absolute (L1) error (people) | Error Distribution Quantiles |            |            |           |          |          |           |           |           |  |
|---|------------------|-----------------------------------|------------------------------|------------|------------|-----------|----------|----------|-----------|-----------|-----------|--|
|   |                  |                                   | q0.005                       | q0.025     | q0.050     | q0.250    | q0.500   | q0.750   | q0.950    | q0.975    | q0.995    |  |
| <b>Bell Simulation (Housing Units [HU] population only)</b> |                  |                                   |                              |            |            |           |          |          |           |           |           |  |
| <b>All blocks with at least 1 housing unit. . . . .</b>     | <b>6,379,963</b> | <b>5.42</b>                       | <b>-47</b>                   | <b>-18</b> | <b>-11</b> | <b>-3</b> | <b>0</b> | <b>1</b> | <b>12</b> | <b>23</b> | <b>73</b> |  |
| All blocks with at least 1 census HU person . . .           | 6,188,078        | 5.58                              | -48                          | -18        | -12        | -3        | 0        | 1        | 13        | 24        | 74        |  |
| Blocks with 1–9 HU persons. . . . .                         | 1,825,130        | 0.63                              | -4                           | -2         | -1         | 0         | 0        | 0        | 2         | 3         | 5         |  |
| Blocks with 10–99 HU persons . . . . .                      | 3,663,268        | 4.39                              | -15                          | -10        | -8         | -3        | -1       | 1        | 12        | 19        | 41        |  |
| Blocks with 100–999 HU persons. . . . .                     | 691,557          | 23.24                             | -97                          | -57        | -42        | -16       | -8       | 6        | 61        | 97        | 222       |  |
| Blocks with 1,000+ HU persons . . . . .                     | 8,123            | 154.72                            | -430                         | -293       | -242       | -135      | -78      | 31       | 388       | 591       | 1,210     |  |
| <b>Schafer Simulation (HU population only)</b>              |                  |                                   |                              |            |            |           |          |          |           |           |           |  |
| <b>All blocks with at least one HU . . . . .</b>            | <b>6,379,963</b> | <b>1.47</b>                       | <b>-10</b>                   | <b>-5</b>  | <b>-4</b>  | <b>-1</b> | <b>0</b> | <b>1</b> | <b>4</b>  | <b>5</b>  | <b>10</b> |  |
| Blocks with 0 HU persons . . . . .                          | 191,885          | 0.10                              | 0                            | 0          | 0          | 0         | 0        | 0        | 1         | 1         | 2         |  |
| Blocks with 1–9 HU persons. . . . .                         | 1,825,130        | 0.42                              | -3                           | -2         | -1         | 0         | 0        | 0        | 1         | 2         | 3         |  |
| Blocks with 10–99 HU persons . . . . .                      | 3,663,268        | 1.53                              | -7                           | -4         | -3         | -1        | 0        | 1        | 3         | 4         | 7         |  |
| Blocks with 100–999 HU persons . . . . .                    | 691,557          | 4.14                              | -18                          | -11        | -9         | -3        | 0        | 3        | 9         | 12        | 19        |  |
| Blocks with 1,000+ HU persons . . . . .                     | 8,123            | 14.48                             | -57                          | -39        | -31        | -11       | 1        | 12       | 30        | 37        | 53        |  |

Source: U.S. Census Bureau, Research and Methodology Directorate.